CAVE: Interpretable 3D Neural Object Volumes for Robust Conceptual Reasoning

Nhi Pham[1], Artur Jesslen[3], Bernt Schiele[1], Adam Kortylewski[2]*, Jonas Fischer[1]*

* Equal senior advisorship

Max Planck Institute for Informatics, Saarland Informatics Campus, Germany[1]

CISPA Helmholtz Center for Information Security, Germany[2]

University of Freiburg, Germany[3]

MAX PLANCK INSTITUTE FOR INFORMATICS

SIC Saarland Informatics Campus

CVPR JUNE 3-7, 2026 DENVER COLORADO

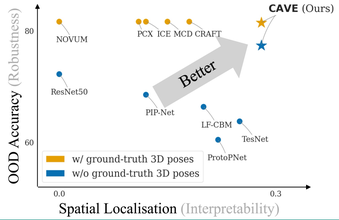
CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

universität freiburg

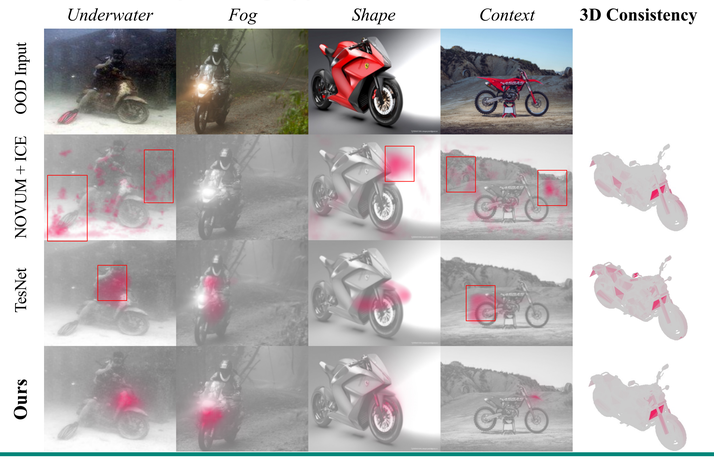
Interpretable and OOD-robust Classification

(a) Explicit 3D-Aware Concept Representation

(b) Robustness vs. Interpretability Trade-off



Motivation: Why existing approaches fall short

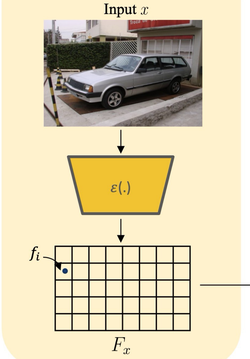


Contributions

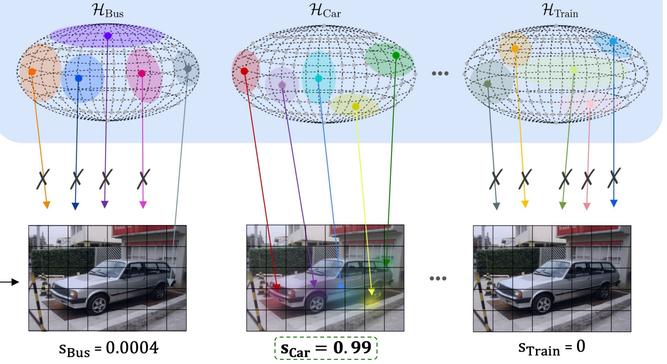
- ✓ **CAVE**: sparse concept for **robust** and **interpretable** classification
- ✓ **NOV-aware attribution** for **faithful** volumetric concepts
- ✓ **3D-C** metric as **part-annotation-free concept consistency** measure
- ✓ **No ground-truth 3D poses needed** at training time

Method: CAVE – Concept-Aware Volumes for Explanations

(a) Extract Image Features

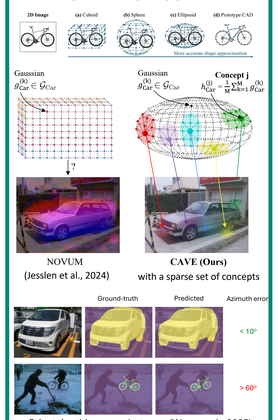


(b) Interpretable Concept-based Neural Object Volumes $\mathcal{H} = [\mathcal{H}_{\text{Bus}}; \mathcal{H}_{\text{Car}}; \ldots; \mathcal{H}_{\text{Train}}]$



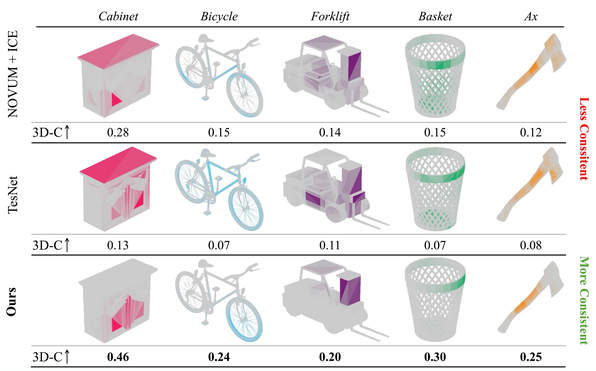
(c) Classify with Concept Matching for each class $s_y = \phi(F_x, \mathcal{H}_y) = \sum_i \max_{j \leq D} f_i \cdot h_y^{(j)}$

Preliminaries



Method: 3D Consistency (3D-C) of Concepts

- ✓ Benchmark **concept consistency** under **OOD** settings
- ✓ An **image-part-annotation-free** metric using **object geometry**



Result: Robust and Interpretable Concepts

1 – OOD-robust classification with sparse 3D-aware concepts

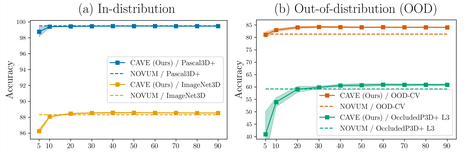
Models	W/o Ground-truth 3D Pose	In-distribution		Out-of-distribution (OOD)	
		Pascal3D+	ImageNet3D	Occluded P3D+	OOD-CV
LF-CBM (Oikarinen et al., 2023)	**Yes**	98.4	83.3	66.4	73.5
ProtoPNet (Chen et al., 2019)	**Yes**	97.4	74.0	60.5	71.2
TesNet (Wang et al., 2021)	**Yes**	97.6	77.9	63.8	70.1
PIP-Net (Nauta et al., 2023a)	**Yes**	95.7	51.0	68.6	60.0
MGProto (Wang et al., 2025a)	**Yes**	97.2	64.2	73.8	72.3
CAVE (Ours)	**Yes**	**99.0** (± 0.03)	**84.6** (± 0.02)	**76.8** (± 0.51)	**80.3** (± 0.27)
CAVE (with full 3D supervision)	No	99.4 (± 0.02)	88.5 (± 0.03)	81.3 (± 0.30)	84.0 (± 0.21)
NOVUM (with full 3D supervision)	No	99.5	88.3	81.7	81.3

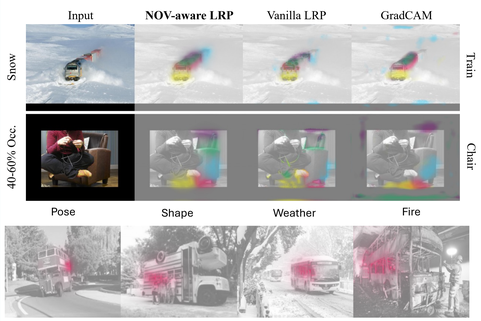
2 – Spatially consistent concepts

	Models	*Localise.* ↑	*Coverage* ↑	*3D Consistency* (3D-C.) ↑			
		Pascal-Part		Pascal3D+	ImageNet3D	OccludedP3D+	OOD-CV
Post-hoc	NOVUM + CRAFT (Fel et al., 2023b)	0.18	0.42	0.28	0.26	0.15	0.15
	NOVUM + MCD (Vielhaben et al., 2023)	0.15	0.34	0.16	0.25	0.11	0.14
	NOVUM + ICE (Zhang et al., 2021)	0.12	0.44	0.28	0.27	0.15	0.15
	NOVUM + PCX (Dreyer et al., 2024)	0.11	0.33	0.10	0.21	0.08	0.11
Ad-hoc	LF-CBM (Oikarinen et al., 2023)	0.20	0.56	0.15	0.14	0.13	0.11
	ProtoPNet (Chen et al., 2019)	0.22	0.43	0.19	0.13	0.21	0.09
	TesNet (Wang et al., 2021)	0.25	0.44	0.20	0.18	0.18	0.12
	PIP-Net (Nauta et al., 2023a)	0.12	0.13	0.09	0.09	0.07	0.00
	MGProto (Wang et al., 2025a)	0.25	0.35	0.19	0.16	0.16	0.07
	CAVE (Ours)	**0.28** (± 0.001)	**0.80** (± 0.002)	**0.40** (± 0.001)	**0.40** (± 0.001)	**0.23** (± 0.006)	**0.24** (± 0.002)
	CAVE (with full 3D supervision)	0.28 (± 0.001)	0.87 (± 0.002)	0.42 (± 0.001)	0.43 (± 0.0003)	0.23 (± 0.010)	0.26 (± 0.001)

(a) In-distribution

(b) Out-of-distribution (OOD)





CAVE